IMPROVING ENGLISH AND CHINESE AD-HOC RETRIEVAL: TIPSTER TEXT PHASE 3 FINAL REPORT

Kui-Lam Kwok

Computer Science Department, Queens College, CUNY
65-30 Kissena Boulevard, Flushing, NY 11367
email: kwok@ir.cs.qc.edu
phone: (718) 997 3482/3500

ABSTRACT

We investigated both English and Chinese ad-hoc information retrieval (IR). Part of our objectives is to study the use of term, phrasal and topical concept level evidence, either individually or in combination, to improve retrieval accuracy. For short queries, we studied five term level techniques that together lead to improvements over standard ad-hoc 2-stage retrieval some 20% to 40% for TREC5 & 6 experiments.

For long queries, we studied linguistic phrases as evidence to re-rank outputs of term level retrieval. It brings small improvements in both TREC5 & 6 experiments, but needs further confirmation. We also investigated clustering of output documents from term level retrieval. Our aim is to separate relevant and irrelevant documents into different clusters, and to re-rank the output list by groups based on query and cluster-profile matching. Investigation is still on-going.

For Chinese IR, many results were confirmed or discovered. For example, accurate word segmentation is not as important as first thought, but short-word segmentation is preferable to long-word (phrase). Simple bigram representation can give very good retrieval. A stopword list is not necessary; and presence of non-content terms does not hurt evaluation results much. One only needs screening out statistical stopwords of high frequency. Character indexing by itself is not competitive, but is useful for augmenting short-words or bigrams. Best results were obtained by combining retrievals of bigram and short-word with character representation. Chinese IR returns better precision than English, and it is not clear if this is a language-related, or collection-related phenomenon.

1. INTRODUCTION

As increasing amounts of computer-readable texts are becoming available on the web or on CDROMs, text searching and detection has become an indispensable tool for information users and analysts of all walks of life. Up till the late 1980's, research in text retrieval has been mainly with small collections of a few thousand items. Since 1990, with the foresight of the TIPSTER and TREC programs, substantial progress has been made to advance the state-of-the-art in text detection and ad-hoc information retrieval (IR) methodologies. Examples include: availability, experimentation and uniform evaluation of gigabyte-size collections, term weighting improvements, 2-stage 'pseudo-feedback' retrieval strategy, recognition of difficulties of short queries versus long, use of phrases, treatment of foreign languages for multilingual retrieval, among others. This investigation builds upon previous findings to bring further advances in this field using our PIRCS system.

We have participated in all past TREC experiments with consistently superior results. Since 1996, we have also participated in the TIPSTER Text Phase 3 program. This report serves to summarize work that has been done, and some of the important findings for both English and Chinese IR. Section 2 and 3 gives an overview of our PIRCS system and the 2-stage retrieval strategy. Section 4 presents our work for English ad-hoc retrieval employing term, phrasal and topical concept levels of evidence. Section 5 describes various Chinese retrieval experiments. Section 6 has the conclusions.

2. PIRCS RETRIEVAL SYSTEM

The software program we use for our Tipster 3 investigations is PIRCS (acronym for Probabilistic Indexing and Retrieval - Components - System). It is a document retrieval system that has been developed in-house since the mid 1980s. It is based on the probabilistic indexing and retrieval approach, conceptualized as a three layer network with adaptive capability to support feedback and query expansion,

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
OCT 1998		**00-00-1998 to 00-00-1998**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Improving English and Chinese Ad-hoc Retrieval: TIPSTER Text Phase 3 Final Report	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Queens College, CUNY,Department of Computer Science,65-30 Kissena Boulevard,Flushing,NY,11367**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998. Sponsored by the Defense Advanced Research Projects Agency.

14. ABSTRACT

We investigated both English and Chinese ad-hoc information retrieval (IR). Part of our objectives is to study the use of term, phrasal and topical concept level evidence, either individually or in combination, to improve retrieval accuracy. For short queries, we studied five term level techniques that together lead to improvements over standard ad-hoc 2-stage retrieval some 20% to 40% for TREC5 & 6 experiments. For long queries, we studied linguistic phrases as evidence to re-rank outputs of term level retrieval. It brings small improvements in both TREC5 & 6 experiments, but needs further confirmation. We also investigated clustering of output documents from term level retrieval. Our aim is to separate relevant and irrelevant documents into different clusters, and to rerank the output list by groups based on query and cluster-profile matching. Investigation is still on-going. For Chinese IR, many results were confirmed or discovered. For example, accurate word segmentation is not as important as first thought, but short-word segmentation is preferable to long-word (phrase). Simple bigram representation can give very good retrieval. A stopword list is not necessary; and presence of non-content terms does not hurt evaluation results much. One only needs screening out statistical stopwords of high frequency. Character indexing by itself is not competitive, but is useful for augmenting short-words or bigrams. Best results were obtained by combining retrievals of bigram and short-word with character representation. Chinese IR retums better precision than English, and it is not clear if this is a language-related, or collection-related phenomenon.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

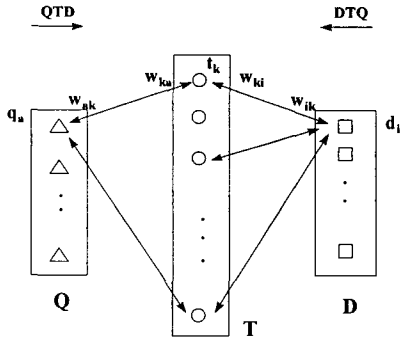

Fig.1a PIRCS's 3 Layer Network for Retrieval

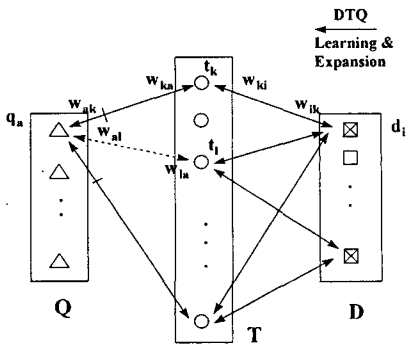

Fig.1b Query-Focused Learning & Expansion

and operates via activation spreading. The network with three levels of query Q, term T and document D nodes are connected with bi-directional weighted edges as shown in Fig.1a for retrieval. Fig.1b shows the network for performing learning where both the edge weights and the architecture can adapt. Learning takes place when some relevant documents are known for a query. The basic model evaluates a retrieval status value (RSV) for each query document pair (q_a d_i) as a combination of a document-focused QTD process that spreads activation from query to document through common terms k, and an analogous query-focused DTQ process operating vice versa, as follows:

$$RSV = \alpha * \Sigma_k \, w_{ik} * S(q_{ak}/L_a) + (1-\alpha) * \Sigma_k \, w_{ak} * S(d_{ik}/L_i)$$

where $0 \le \alpha \le 1$ is a combination parameter for the two processes, q_{ak} and d_{ik} are the frequency of term k in a query or document respectively, L_a, L_i are the query or document lengths, and S(.) is a sigmoid-like function to suppress outlying values. A major difference of our model from other probabilistic approaches is to treat a document or query as non-monolithic, but constituted of conceptual components (which we approximate as terms). This leads us to formulate in a collection of components rather than documents, and allows us to account for the non-binary occurrence of terms in items in a natural way. For example, in the usual discriminatory weighting formula for query term k: w_{ak} = log [p*(1-q)/(1-p)/q], p = Pr(term k present | relevant) is set to a query 'self-learn' value of q_{ak}/L_a based on the assumption that a query is relevant to itself, and q = Pr(term k present | ~relevant) is set to F_k/M, the collection term frequency of k, F_k, divided by the total number of terms M used in the collection. This we call the inverse collection term frequency ICTF. It differs from the usual inverse document frequency IDF in that the latter counts only the presence and absence of terms in a document, ignoring the within-document term frequency. Moreover, as the system learns from relevant documents, p can be trained to a value intermediate between the basic self-learn value and that given by the known relevants according to a learning procedure [1]. Our system also uses two-word adjacency phrases as terms to improve on the basic single word representation. Documents of many thousands or more words long can have adverse effect on retrieval. PIRCS deals with the problem by simply segmenting long documents into approximately equal sub-documents of 550-word size and ending on a paragraph boundary. For the final retrieval list, retrieval status values (RSV) of the top three sub-documents of the same document are combined with decreasing weights to return a final RSV. This in effect favors retrieval of longer documents that contain positive evidence in different sub-parts of it. PIRCS has participated in all previous TREC 1-6 blind retrieval experiments and consistently returned some of the best results, see for example [2].

3. TWO-STAGE AD-HOC STRATEGY

Automatic ad-hoc retrieval refers to the environment where a user attempts to retrieve relevant documents from an existing collection by issuing 'any' query. We have experimented only with natural language queries that are derived from TREC topics. It is a difficult problem because the query wordings are unknown beforehand, and its topical content is unpredictable. Moreover, there will not be any example relevant documents that a system can rely on for training purposes like in a routing situation.

To improve the accuracy of ad-hoc retrieval, it is now a common practice to adopt a 2-stage retrieval strategy. Under the right circumstances this can give substantial improvements over single stage. In a 1-

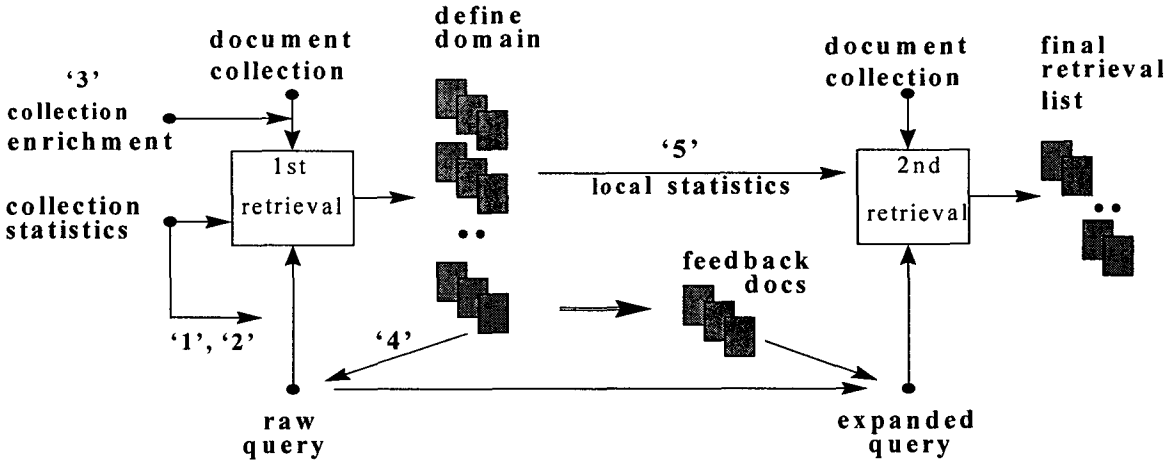

Fig.2 Two-Stage Retrieval and Methods of Improvements

stage retrieval, the raw query which is a user-provided description of information needs is directly employed by the retrieval algorithm to assign a retrieval status value (RSV) to each document in a collection, and the ranked list of documents is interpreted as the final retrieval result. In a 2-stage strategy, this initial ranked list is interpreted as but an intermediate step. The set of n top-ranked documents of the initial retrieval is assumed relevant, even though the user has not made any judgment. These 'pseudo-relevant' documents are then used to modify the weight of the initial query according to some learning procedure, as well as to expand the query with terms from these documents based on some selection criteria like frequency of occurrence. The modified query is then used to do a second retrieval, and the resultant ranked list becomes the final result. This helps because if the raw query is reasonable and the retrieval engine is any good, the initial top n documents can be considered as defining the topical domain of the user need and should have a reasonable density of relevant or highly related documents, and the procedure simulates real relevance feedback.

Traditionally, real relevance feedback can give very large improvements in average precision, like 50 to over 100%. Experiments with our PIRCS system have shown that this 2-stage of ad-hoc method works more often than not, about 2 out 3 times (35 queries in TREC-5 and 32 in TREC-6 out of 50 queries each), and the average precision for a set of queries can improve a few to over 20%. The process of a 2-stage retrieval is depicted in Fig.2.

In all of our work, this 2-stage approach is used in our retrieval experiments. Some tables below show initial 1-stage results for comparison purposes.

4. ENGLISH AD-HOC RETRIEVAL

An important finding in the TREC experiments is that short queries have substantially different retrieval properties from long ones. We consider short queries as those with a few content terms and are popular in casual environments such as web searching. Serious users wanting more exhaustive and accurate searching should issue longer paragraph-size queries with some related conceptual terms. They usually return better effectiveness because longer exposition of needs can reduce the ambiguity problem due to homographs and the descriptive deficiency due to synonyms. The 2-stage retrieval approach has been shown in several years of TREC experiments to improve over 1-stage for both query types. Our work has investigated additional methods to enhance retrieval accuracy for this strategy.

4.1 Term Level Evidence

We studied several methods for improving our approach of 2-stage pseudo-relevance feedback retrieval for short queries [3]. These are related to using single term statistics and evidence, and include (see Fig.2): 1) avtf query term weighting, 2) variable

high frequency Zipfian threshold, 3) collection enrichment, 4) enhancing term variety in raw queries, and 5) using retrieved document local term statistics. Avtf employs collection statistics to weight terms in short queries [4] where term importance indication is generally not available. Variable high frequency threshold defines statistical stopwords based on query length. Collection enrichment adds external collections to the target collection under investigation so as to improve the chance of ranking more relevant documents in the top n for the pseudo-feedback process. Adding term variety to raw queries means adding highly associated terms from the domain-related top n documents based on mutual information values. Making the query longer may improve 1st stage retrieval. And retrieved document local statistics re-weight terms in the 2nd stage using the set of domain-related documents rather than the whole collection as used during the initial stage. Results using these methods are tabulated in Table 1 where we show some of the popular evaluation measures: RR - the number of relevant documents returned after retrieving 1000 documents; AvPre - the non-interpolated average precision; P@10 - the precision at 10 documents retrieved, and R.Pre - the recall precision at the point where the number retrieved is exactly equal to the number of relevant documents

It can be seen that standard 2-stage strategy performs about 9% to 15% better than initial retrieval using the AvPre measure as reference (TREC5 .161 vs. .140, TREC6 .240 vs. .220). The other techniques successively bring further improvements, accumulating to about 20 to 40% over the standard 2nd stage retrieval results (TREC5 .239 vs. .161, TREC6 .289 vs. .240).

It is found that collection enrichment also works for long queries. It is an attractive technique since searchable texts are increasingly available nowadays.

	1st Retr	2nd Retr	Avtf	Var. Thld	Coll. Enrich	M.I. Terms
	←TREC5 50 Short Queries →					
RR	1763	2279	2335	2635	2732	2787
AvPre	**.140**	**.161**	.181	.214	.234	**.239**
P@10	.290	.284	.326	.372	.382	.404
R.Pre	.179	.191	.210	.249	.270	.271
	←TREC6 50 Short Queries →					
RR	2188	2272	2384	2517	2656	2738
AvPre	**.220**	**.240**	.258	.258	.284	**.289**
P@10	.334	.372	.402	.388	.444	.442
R.Pre	.262	.264	.291	.287	.312	.311

Table1: Term Level Retrieval Enhancement

We envisage that so long as the external text falls within similar topical domain of the query, it could be helpful as an enrichment tool. It goes quite a way to improve the accuracy of retrieval, especially in the difficult ad-hoc, short query situations.

4.2 Phrase Level Evidence

Investigators in IR are aware of the simplistic and inadequate representation of document content based on a bag of single word stems or some 2-word adjacency phrases. To a certain extent this is dictated by the requirements that text retrieval systems have to support large scale environments as well as unpredictable, diverse needs. Many previous attempts, including Tipster contractors (e.g. [5]), have been made to include more sophisticated phrasal representation in order to improve retrieval results. They have not worked as well as content terms or generally been inconclusive.

We also investigated phrasal evidence for retrieval, but only to the extent that it is used to refine results that have been obtained via term level retrieval. Only long queries are considered since queries with too few phrases would not provide sufficient evidence to work with. Specifically, we use phrasal evidence to re-rank a retrieved document list so as to promote more relevant documents earlier in the list. This could lead to higher density of true relevant documents in the 1st stage retrieval, thereby improving 'pseudo-feedback' for the 2nd stage downstream. The 2nd stage retrieval list could similarly be re-ranked to return better effectiveness as well.

A query is processed into variable length noun phrases using a POS-tagger from Mitre and simple bracketing. (We have also experimented with the BBN tagger before). Given a retrieved document, each noun phrase concept of the query is then matched within up to a 3-sentence context anywhere in the document. When there are matches of two or more terms, appropriate weights are noted for this phrase and the sentence counted. In addition, the amount of coverage of all the query phrases by the document is also a factor by which the original RSV of a document is boosted. However, not all documents have their RSV modified. They need to pass a threshold for coverage.

After many experiments for the TREC 5 and 6 long query environments, the attempt was moderately successful as shown in Table 2. For TREC5, an improvement in AvPre of 4% (.273 vs. .262) was obtained, but in TREC6 only about 1% (.308 vs. .305).

	1st Retr	2nd Retr	2nd Retr Enrich	← Re-rank C	Phrase Rerank A; then 2nd Retr	→ Re-rank C'
	(A)	(B)	(C)	(D)	(C')	(D')
	←TREC5 50 Long Queries →					
RR	2463	3077	3034	3049	3052	3072
AvPre	.220	**.253**	**.262**	.265	.270	**.273**
P@10	.404	.414	.438	.440	.446	.444
R.Pre	.258	.277	.292	.292	.295	.296
	←TREC6 50 Long Queries →					
RR	2537	2947	3043	3064	3074	3088
AvPre	.237	**.264**	**.305**	.310	.304	**.308**
P@10	.402	.452	.492	.498	.488	.490
R.Pre	.278	.296	.326	.332	.327	.331

Table2: Phrase Level Re-ranking Results

More studies need to be done to confirm its utility. Also shown in Table 2 is 2nd stage retrieval without and with collection enrichment (columns B and C). It is seen that this strategy works for long queries too.

4.3 Topical Concept Level Evidence

We have also investigated re-ranking of term level results based on clustering of the retrieval output. The idea is that it is often the case documents are ranked high by matching a query with terms that are related to different unwanted sub-topics or have different senses from those used in the query. Examples of the latter are 'bank', 'deposit' in the money sense, or their river sense. Other terms may disambiguate the true sense in a document, but they may not be present or sufficiently matched to the query. Assuming there are sufficient number of retrieved documents using the terms in their different senses or for different sub-topics, one could separate them into groups by clustering the list. Each group will be characterized by a profile consisting of terms with the highest occurrence frequency within each group. The query can now be matched with the profiles as if they were documents, and the highest ranked profile group would be promoted in ranking.

Because cluster profiles would be important for a query to pick the groups correctly, we have implemented a clustering algorithm that emphasizes on profile forming rather than the more common similarity-matrix based methods such as the single-link or average-link. It is based on the iterative clustering approach of [6,7]. Each sub-document of a (100) top-ranked retrieval list, if not too long or too short, is used as a seed to form a cluster by picking highly similar documents that are not yet clustered. The profile from the resulting group is further iterated until there is no or little change in the profile. Each unclustered sub-document is tested as a seed to form a group, but many failed because fairly stringent conditions need to be satisfied. After the process, there often would be left with sub-documents that belong to no clusters. They are lumped together as 'miscellaneous' and has its profile formed. In a number of queries, this 'miscellaneous' cluster actually contain the most relevant documents. This is the case because there is not sufficient relevant documents to satisfy the group forming criteria, or that their usage of terms are too diverse and non-overlapping.

So far the attempt has not been successful. Several difficulties are noted: the clustering algorithm sometimes does not work well in separating relevant and irrelevant documents into different clusters; often the query may not pick the right cluster to re-rank; and even if the right cluster has been picked, the relevant documents may not rank sufficiently high within the cluster so that a lower AvPre measure may result. The investigation is still ongoing.

5. CHINESE AD-HOC RETRIEVAL

Our research continues the work of other investigators on Chinese IR during Tipster 1&2 (e.g. [8]). We have augmented our PIRCS system to handle the 2-byte encoding of Chinese characters according to the GB2312 convention. During processing, our system can handle both English and Chinese present simultaneously in documents and queries.

5.1 Word Segmentation

A major difference of Chinese writing from English is that a Chinese sentence (which can usually be recognized by a punctuation ending) consists of a continuous string of characters and there is no white-space to delimit words. Words can be one, two or more characters long. At the time, we believed that word segmentation is important for effective Chinese IR. Since efficient word segmentation software for large collections were not available, we relied on an approximate short-word segmenter that was developed by ourselves in house (Queens segmenter [9]). Because the segmenter may not be sufficiently accurate, we actually use characters in addition to short-words for both query and document representation. Earlier work has used word segmentation on queries only and rely on character representation for documents with operators to

combine characters for matching query words [8].

The blind Chinese retrieval results in both TREC 5 and 6 showed that our short-word plus character indexing method works very well, since we have returned the best automatic retrieval evaluations for both years [10,11]. It also demonstrates that the PIRCS retrieval model can handle both English and Chinese languages equally good. After the blind TREC5 experiment, we further optimize parameters in PIRCS such as sub-document size, number of documents and number of terms to use for 2nd stage retrieval to obtain better results [12] as shown in Table 3.

It can be seen that two-stage retrieval is good for both English and Chinese, leading to improvements in AvPre of some 15% to 31% (.452 vs. .392 and .384 vs. .293) over initial 1st stage retrieval. Moreover, long queries perform better than short ones as in English, between 17% and 22% (.452 vs. .384 and .603 vs. .476). These Chinese queries return surprisingly good results even though the segmentation is approximate. It is not clear if the language characteristics itself may be a factor contributing to this.

5.2 Comparing Segmenters

Word segmentation is a big issue for Chinese since linguistics-strong applications such as POS tagging, sentence parsing, machine translation, text to voice, etc. are all dependent on words being accurately identified to do well. It would therefore be interesting to see if better word segmentation could lead to more accurate retrieval.

	TREC5			
	← 28 Long Queries →		← 28 Short Queries →	
	1st Stage	2nd Stage	1st Stage	2nd Stage
RR	1944	2015	1615	1707
AvPre	**.392**	**.452**	**.293**	**.384**
P@10	.546	.600	.389	.511
R.Pre	.403	.452	.316	.389

	TREC6			
	← 26 Long Queries →		← 26 Short Queries →	
	1st Stage	2nd Stage	1st Stage	2nd Stage
RR	2738	2791	2277	2547
AvPre	**.551**	**.603**	**.376**	**.476**
P@10	.808	.869	.615	.712
R.Pre	.532	.567	.401	.463

Table3: 1st and 2nd Stage Chinese Retrieval Results

We have done manual analysis of our approximate segmenter for correctness using the 54 TREC 5 & 6 topics and concluded that its recall and precision measures for segmenting sentences into short-words are about mid to high 80%. These figures are approximate because even native speakers sometimes disagree on the correct segmentation. We have also analyzed a segmenter from UMASS [13] that is based on a unigram model. It can be trained from a collection that has been segmented based on a lexicon list. It segments a sentence by evaluating possible choices and selecting the one with the highest probability of the trained model. Our opinion is that its recall and precision values vary between about 90% to low-90%, approximately 5% better than ours. We used both segmenters to investigate the Chinese collection and did retrieval using our PIRCS system under the same parameter settings. The result is presented in Table 4 below. In this table, TREC5 precision values took account of larger lexicons (Section 5.3) and are better than those in Table 3.

It is a bit surprising to see that results using the two segmenters are very similar. It appears that better segmentation may not mean better retrieval. It is possible that these two segmenters are not sufficiently different to reflect any significant changes in results. A very high quality segmenter of 95% or higher accuracy may tell a different story.

5.3 Lexicon Size Effects

We made further studies of retrieval using our approximate segmenter to see how it might depend on the lexicon used. Our segmentation procedure depends on some simple, approximate language usage rules as well as an initial lexicon list. If a string of Chinese characters is not found on the lexicon, the rules operate

	TREC5			
	← 28 Long Queries →		← 28 Short Queries →	
	Queens	UMASS	Queens	UMASS
RR	2059	2070	1972	1991
AvPre	**.467**	**.460**	**.417**	**.414**
P@10	.625	.589	.554	.561
R.Pre	.471	.453	.413	.412

	TREC6			
	← 26 Long Queries →		← 26 Short Queries →	
	Queens	UMASS	Queens	UMASS
RR	2791	2761	2547	2488
AvPre	**.603**	**.587**	**.476**	**.491**
P@10	.869	.850	.712	.750
R.Pre	.567	.557	.463	.476

Table4: Comparing Queens & UMASS Segmenters

to segment the string into short-words, thereby also discovering unknown words. Our initial lexicon L0 is manually prepared and about 2K in size, minuscule compared to lists used by other investigators for segmentation purposes. By bootstrapping, a larger lexicon list L01 (about 15K) was derived automatically, and it can be used in place of the initial lexicon list for a more refined segmentation.

If a larger initial lexicon list is used, there should be more matching between a document string and the lexicon entries, the approximate rules would be used less often and the resultant segmentation could be more accurate. This would also be true for the derived lexicon. Better segmentation might also affect retrieval favorably.

We have additionally prepared a much larger initial lexicon list L1 (~27K) based on the association list in the Cxterm software. Together with the derived lexicon L11 (43K), we have studied the effects of using these four lexicons for segmentation and retrieval. The results are shown in Table 5. We observe that larger lexicon list can lead to incrementally better AvPre values (.463 vs .455 for long queries and .409 vs .398 for short), but the rate of increase is very slow. The initial 2K lexicon gives surprisingly good results.

	TREC5			
	← 28 Long Queries →		← 28 Short Queries →	
	L0	**L11**	**L0**	**L11**
RR	2059	2061	1958	1975
AvPre	**.455**	**.463**	**.398**	**.409**
P@10	.596	.604	.534	.564
R.Pre	.455	.461	.403	.398

Table5: Lexicon Size Effects on Chinese Retrieval

5.4 Stopword Effects

Stopwords are function words that do not carry much content by themselves, and are usually removed based on a compiled stopword list to improve precision and efficiency. In addition, high frequency terms in a collection, which we call statistical stopwords, are also removed because they are too widespread. On the other hand, stopword removal always carry the risk that one might delete some words that might be crucial for particular queries or documents but in general not very useful. Examples (in English) are words like 'hope' in 'Hope Project' [9], or 'begin' in 'Prime Minister Begin'. They can normally be regarded as not content-bearing, but in the examples given they become crucial. Removing them will adversely affect results. Experiments with and without stopword removal (from a list) however shows that retrieval results are minimally affected. Chinese IR seems to tolerate noisy indexing well. The lesson is not to use any stopword list at all else one might run into perils as discussed. Statistical stopwords are still removed.

5.5 Bigram Representation

We have further experimented with using simpler representation methods such as single characters and bigrams (consecutive overlapping two character) for retrieval. Bigram representation does not need any segmentation or linguistic rules, but often over-generates a large number of indexing terms that are not meaningful to humans. Character indexing is even simpler, but they are highly ambiguous since there are only 6763 distinct characters in the GB2312 scheme. Surprisingly results with single characters are good, though not competitive; but bigram results can rival those of short-words when the queries are long. This has important ramifications since it means that for effective Chinese IR, one need not worry about which segmentation method to use. (More intensive linguistic processing of course still requires accurate segmentation.) For large-scale collections, bigram segmentation is also more efficient time-wise, although it is more expensive space-wise. Table 6 shows examples of retrieval measures using character and bigram representation.

	TREC5			
	← 28 Long Queries →		← 28 Short Queries →	
	Char	**Bigram**	**Char**	**Bigram**
RR	2007	2128	1757	1971
AvPre	**.381**	**.457**	**.318**	**.372**
P@10	.539	.618	.421	.479
R.Pre	.403	.459	.351	.382

	TREC6			
	← 26 Long Queries →		← 26 Short Queries →	
	Char	**Bigram**	**Char**	**Bigram**
RR	2612	2735	2304	2342
AvPre	**.512**	**.574**	**.432**	**.459**
P@10	.785	.827	.723	.658
R.Pre	.507	.547	.433	.460

Table6: Character and Bigram Retrieval Results

5.6 Combining Representations

Since short-word with character and bigram representations separately returns comparable good results, this leads us to investigate whether they can perhaps reinforce each other. Short-words provide effective term matching between a query and a document, but one might have wrong segmentations. Bigrams however are exhaustive and can remedy the situation. Given a collection, we index it both ways. For each query we also index it both ways and perform separate retrievals. Their retrieval lists are then combined based on the RSV of each document i as follows (with α=1/2):

$$RSV_i = \alpha * RSV_{i1} + (1- \alpha) * RSV_{i2}$$

The result, shown in Table 7 as 'sw.c+bi' column, was a further improvement of about 2 to 4% compared with the best of the two base precision without combination for both short and long queries. The price to pay is the doubling of time and space. If for some applications the last bit of effectiveness is important, this is a viable approach. Moreover, this strategy could be realized by having both retrievals performed in parallel on separate hardware, thus without affecting the time of retrieval too much.

Included in Table 7 as the 'bi.c' column is the result of adding characters to bigram indexing, just like adding characters to short-words. Compared to Table 6, it is seen that this is also useful in 3 out of 4 cases, varying from -0.7% (.454 vs.0.457) to +13% (0.489 vs. .432) changes in AvPre for bigram results. Characters are highly ambiguous as indexing terms but there are also Chinese words that are truly single character, and using bigrams only would not lead to correct term matching.

	TREC5			
	← 28 Long Queries →		←← 28 Short Queries →	
	bi.c	**sw.c+bi**	**bi.c**	**sw.c+bi**
RR	2126	2111	1981	1985
AvPre	**.454**	**.471**	**.387**	**.425**
P@10	.600	.621	.511	.539
R.Pre	.456	.468	.405	.423

	TREC6			
	← 26 Long Queries →		←← 26 Short Queries →	
	bi.c	**sw.c+bi**	**bi.c**	**sw.c+bi**
RR	2806	2784	2521	2611
AvPre	**.627**	**.633**	**.489**	**.514**
P@10	.858	.869	.739	.750
R.Pre	.575	.582	.482	.496

Table7: Combining Representations for Retrieval

5.7 Collection Enrichment for Chinese IR

In Section 4.1, we observed that collection enrichment is an effective strategy to improve English ad-hoc retrieval, especially for short queries. Here, we study if this is also true for Chinese.

The TREC5 Chinese collection came from two sources: 24,988 documents from XinHua News Agency (xh) and 139,801 from Peoples' Daily newspaper (pd). In PIRCS, they were segmented into sub-documents of 38,287 and 193,240 items respectively. We use the combined TREC5 and 6 queries numbering 54, and do retrieval with the xh collection as the target but enriched with pd, and vice versa. Some queries do not have any relevants in one of the sub-collections and the actual number of queries for evaluation is less. This is done for the both long and short (title only) versions of the queries. Results are tabulated in Table 8.

It is seen that, except for long queries retrieving on pd and enriched with xh where the AvPre practically remains unchanged (.499 vs. .500), the other cases have improvements of between 3 to 4% over the standard 2nd retrieval without enrichment. The latter already has quite high effectiveness in these cases. Thus, we may say that collection enrichment also works in Chinese.

	Target: xh (enriched by pd)					
	← 52 Long Queries →			←← 52 Short Queries →		
	1st Retr	**2nd Retr**	**2nd Retr enrich**	**1st Retr**	**2nd Retr**	**2nd Retr enrich**
RR	1685	1704	1706	1497	1586	1592
AvPre	.472	**.533**	**.550**	.384	**.445**	**.462**
P@10	.498	.575	.585	.423	.494	.504
R.Pre	.468	.508	.515	.399	.430	.443

	Target: pd (enriched by xh)					
	← 53 Long Queries →			←← 53 Short Queries →		
	1st Retr	**2nd Retr**	**2nd Retr enrich**	**1st Retr**	**2nd Retr**	**2nd Retr enrich**
RR	3174	3264	3269	2763	3066	3052
AvPre	.443	**.500**	**.499**	.325	**.404**	**.416**
P@10	.615	.664	.677	.468	.542	.581
R.Pre	.447	.485	.486	.345	.413	.420

Table 8: Chinese Collection Enrichment Results

6. CONCLUSION

A 2-stage retrieval strategy with pseudo-feedback often returns better ad-hoc results than 1-stage alone. We have further investigated term, phrasal and topical concept level evidence methods for improving retrieval accuracy in this situation. We showed that five term level methods together are effective for enhancing ad-hoc short query results some 20 to 40% for TREC5 & 6 experiments. A particularly useful technique is collection enrichment, which simply adds domain-related external collections to a target collection to help improve 2nd stage retrieval downstream. It brings substantial improvements in many cases and does not hurt much in others. It works for long and short queries in both English and Chinese IR.

With long queries we showed that using linguistic phrases to match within document windows as further evidence to re-rank retrieval output can lead to some small improvements. We also studied re-ranking of output documents based on topical concept level evidence using document clustering, but the effort has so far not been successful.

Contrary to expectations, word segmentation is not crucial for Chinese IR. Simple bigrams or short-word with character indexing can produce very good results. A manual stoplist is also unnecessary; one only needs to screen out high frequency statistical stopwords. Best results are obtained by combining retrievals using multiple representations.

For the future, it will be interesting to see if phrasal evidence can be employed for Chinese IR, and to study how to improve its usefulness. Topical clustering for enhancing retrieval, display and for data reduction in general are also important issues for large scale IR.

ACKNOWLEDGMENTS

This research is partially supported by a contract from the U.S. Department of Defense MDA904-96-C-1481. I like to express my appreciation to R. Weischedel for use of the BBN POS tagger; L. Hirschman for the Mitre POS tagger and W.B. Croft for the UMASS Chinese segmenter.

REFERENCES

[1] Kwok, K.L. "A network approach to probabilistic information retrieval". ACM Transactions on Office Information System, 13:324-353, July 1995.

[2] Voorhees, E. & Harman, D. "Overview of the Sixth Text REtrieval Conference (TREC 6). In: The Sixth Text REtrieval Conference (TREC-6), E. Voorhees & D.K. Harman (eds.) NIST Special Publication 500-240, Gaithersburg, MD 20899. pp.1-24, 1998.

[3] Kwok, K.L. & M. Chan. "Improving two-stage ad-hoc retrieval for short queries." in Proc. 21st Ann. Intl. ACM SIGIR Conf. on R&D in IR. pp.250-6, 1998.

[4] Kwok, K.L. "A new method of weighting query terms for ad-hoc retrieval". Proc. 19th Annual Intl. ACM SIGIR Conf. on R&D in IR. ETH, Zurich, Aug. 18-22, 96. pp.187-195, 1996.

[5] Strzalkowski, T. "Natural language information retrieval: Tipster-2 final report". Proc. of Tipster Text Program (Phase 2). pp.143-8, Sept., 1996.

[6] Rocchio, J.J. Jr. "Document retrieval systems - optimization and evaluation" Ph.D. thesis, Harvard University [1966].

[7] Schiminovich, S. "Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm". Info. Stor. & Retr. 6:417-435, 1971.

[8] Boisen, S., Crystal, M., Petersen, E., Weischedel, R., Broglio, J., Callan, J., Croft, B., Hand, T., Keenan, T., Okurowski, M. "Chinese information extraction & retrieval". Proc. of Tipster Text Program (Phase 2). pp.109-119, Sept., 1996.

[9] Kwok, K.L. "Lexicon effects on Chinese information retrieval". Proc. of 2nd Conf. on Empirical Methods in NLP. Cardie, C. & Weischedel, R. (eds). Brown Univ., Aug.1-2, 1997. pp.141-148.

[10] Kwok, K.L. & Grunfeld, L. "TREC-5 English and Chinese retrieval experiments using PIRCS". In: Information Technology: The Fifth Text REtrieval Conference (TREC-5), E.M. Voorhees & D.K. Harman, eds. NIST Special Publication 500-238, Gaithersburg, MD 20899. pp.133-142, 1997.

[11] Kwok, K.L., Grunfeld, L. & Xu, J.H. "TREC-6 English and Chinese retrieval experiments using PIRCS". In: The Sixth Text REtrieval Conference (TREC-6), E. Voorhees & D.K. Harman, eds. NIST Special Publication 500-240, Gaithersburg, MD 20899. pp.207-214 1998.

[12] Kwok, K.L. "Comparing representations for Chinese information retrieval". Proc. 20th Annual Intl. ACM SIGIR Conf. on R&D in IR. Philadelphia, Jul 27-31, 1997. pp.34-41.

[13] Ponte, J. & Croft, W.B. "Useg: a retargetable word segmentation procedure for information retrieval". In: Symposium on Document Analysis & Information Retrieval (SDAIR 1996)